Chapter 5

Lake Malawi genetic diversity

5.1 Introduction

The Lake Malawi radiation has generated over 500 species in less than five million years, involving divergence in habitat use, size, shape, diet, feeding apparatus, sex determination, male breeding colour, and many other traits [83]. These phenomena present an opportunity to observe hundreds of varied, recent, and in some cases ongoing, speciation events at different stages along the speciation continuum, ranging from newly forming varieties to pairs of well defined species with very strong or complete reproductive isolation. Studying hundreds of 'snapshots' of speciation at different stages along the continuum, all in a similar genetic background, will give us unprecedented insight into the genomic patterns of divergence that underlie the build-up of reproductive isolation this system. We should be able to see whether the genomic patterns are consistent or vary between sub-groups of species, and also if there are genes or genomic regions that tend to be involved repeatedly and thus qualify as 'speciation genes' in cichlids.

Genomic divergence continues to accumulate after speciation is complete, both due to neutral processes and due to continued adaptive evolution in the newly formed species. In fact, reproductive isolation may not be sufficient for achieving a speciation event of lasting effect; it is a widely held view that in order for species to co-exist, they must achieve sufficient ecological differentiation to reduce direct competition [168, 169, 170, 171] (for an opposing view and evidence see [172]). Therefore, post-speciation adaptive evolution may be as important as pre-speciation divergence for generating and maintaining species diversity, and can be observed in Lake Malawi by comparing older fully-isolated species.

Overall, Lake Malawi cichlids promise to provide a rich picture of genome evolution on a timescale that bridges the gap between micro- and macro-evolutionary studies, with a particular focus on genetic basis of functional diversification. Thus, this study contributes to our understanding of how vertebrate genomes evolve and function and the long term impact of knowledge of genome function includes implications for animal and human health.

Previous studies demonstrate the utility of Lake Malawi cichlids in addressing major questions in evolutionary ecology and genomics, but only scratch the surface, focussing on a limited number of species and/or genes. The progress of research in the large cichlid radiations of lakes Malawi, Tanganyika, and Victoria has been hampered by difficulties in identifying species relationships, in reconstructing past geographical situations, and in controlling for possible introgression from non-sister taxa [86]. Therefore, a thorough characterisation the genomic diversity of a complete large adaptive radiation, and reconstructing of its evolutionary history, including the relative timing, frequency and sequence of evolution of major adaptive innovations are fundamental steps for making this complex and fascinating system tractable for in-depth investigation.

Here I describe initial analysis of the Lake Malawi samples collected in Spring 2013 and Autumn 2014 as listed in chapter 3. This will be extended by adding further data from Summer 2015 and additional analysis to form the basis of a future publication.

5.2 Genomic diversity

Variant calling for the Lake Malawi samples (as defined in section 3.2.3) against the *Metriaclima zebra* reference genome, from which divergence was 0.2-0.3% (0.9% for *A. rujewa*), resulted (after filtering) in 20,673,877 SNPs and 2,859,560 short insertions and deletions (1.2 - 1.8 million variants per individual, except the outgroup *A. rujewa* with 5.7 million variants; Figure 5.1).

I used the frequency of heterozygous sites as a simple summary statistic to estimate within-species nucleotide site diversity (π). Heterozygosity is indicative of long-term effective population size (N_e) over the past of order N_e generations [173]. As shown in Figure 5.2, π in all Lake Malawi species is within a relatively narrow range between ~1/2,000 and ~1/700 (except for the *A. calliptera* sample from Kitai Dam which has a much lower π estimate of ~1/12,000, presumably due to a strong population bottleneck/founder effect).

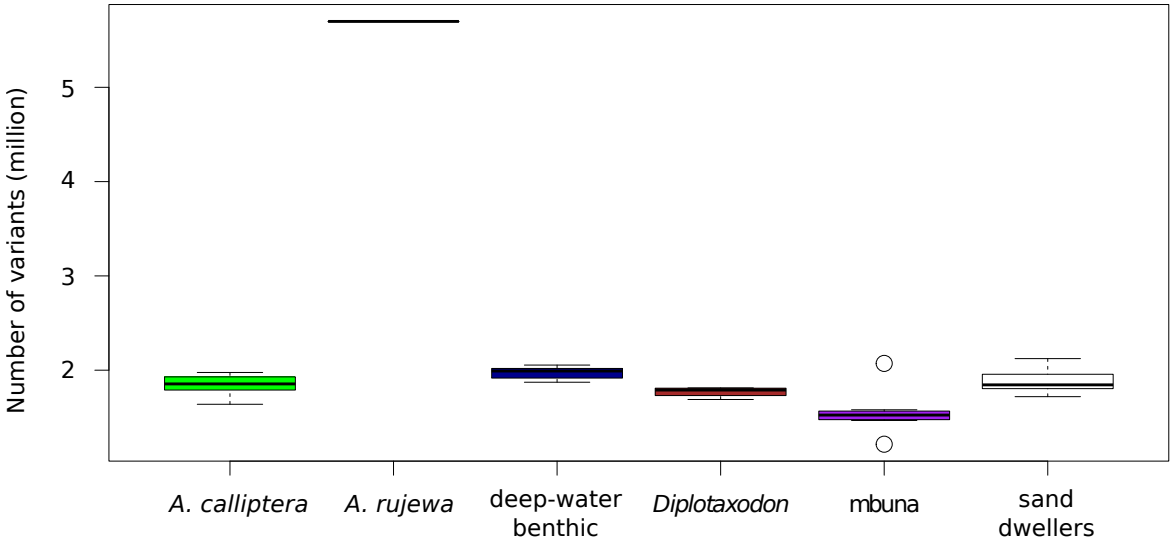

Fig. 5.1 **Variants called against *M. zebra* genome in Lake Malawi samples.**

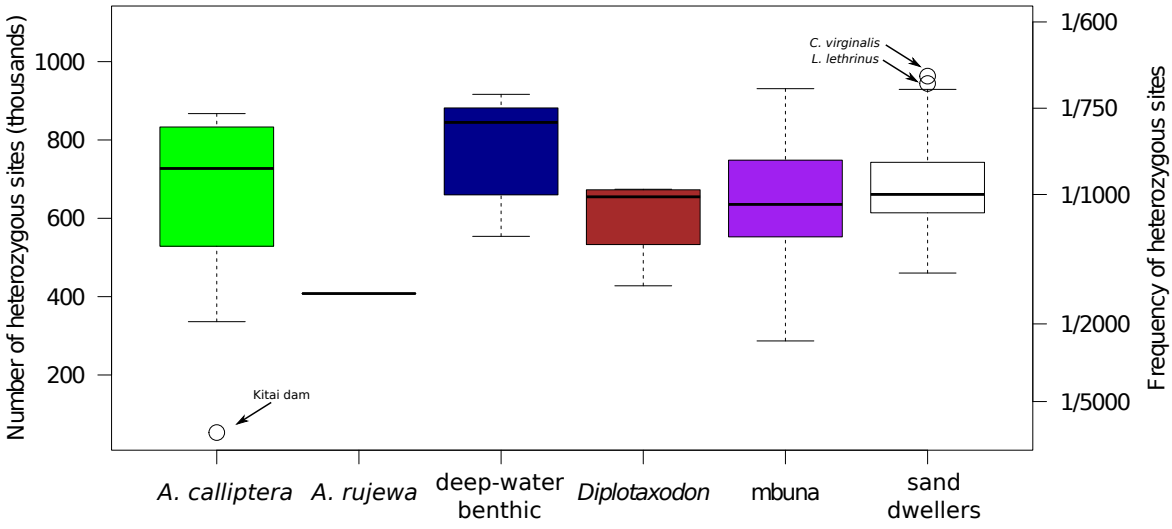

Fig. 5.2 **Frequency of heterozygous sites in Lake Malawi samples.**

The relationship between π and N_e in an ideal population (as defined in section 1.1.3) is:

$$\pi = 4N_e\mu \tag{5.1}$$

where μ is the per-generation mutation rate.

A direct estimate of μ in cichlids is not available. However, if we make the assumption that cichlid μ is similar to human ($\mu \approx 1.5 \times 10^{-8}$ [15]), we can obtain long term N_e estimates between ~7,200 and ~24,300 (~1,300 for the *A. calliptera* sample from Kitai Dam).

To complement measures of within-species genetic diversity (π), I also estimated between-species divergence by calculating the F_{ST} statistic for all pairs of species. The results, shown in Figure 5.3, revealed that:

1. *Diplotaxodon macrops* and *Diplotaxodon* 'macrops black dorsal', a putative new species studied by Genner *et al.* [128], are genetically virtually undistinguishable, at least in terms of genome-wide average F_{ST}. It is possible that their genetic differentiation is limited to a very small number highly diverged genomic loci, as seen for example between German carrion and Swedish hooded crows [77].
2. F_{ST} between the putative outgroup species *A. rujewa* and the rest of the Lake Malawi species varies from 0.629 to 0.857, suggesting that some shared variation remains between *A. rujewa* and Lake Malawi, despite the relatively high level of divergence.
3. Excluding the two special cases above, F_{ST} between species within the Lake Malawi sample set varies between 0.036 to 0.661. The lower value is between *Diplotaxodon limnothrissa* and *Diplotaxodon macrops*, and similar values are also estimated for other pairs of species (e.g. *Copadichromis virginalis* vs. *Copadichromis quadrimaculatus* and *Ctenopharynx intermedius* vs. *Ctenopharynx nitidus*). Such F_{ST} levels are virtually the same as between the two ecomorphs of Lake Massoko described in chapter 6, suggesting that for example the divergence between the two sympatrically occurring *Diplotaxodon* species could be studied with a similar approach as applied for Massoko. At 0.661, the highest F_{ST} estimate is similar to divergence between two wild and phenotypically virtually undistinguishable zebrafish strains from southern India and Bangladesh (F_{ST}=0.64) [174].

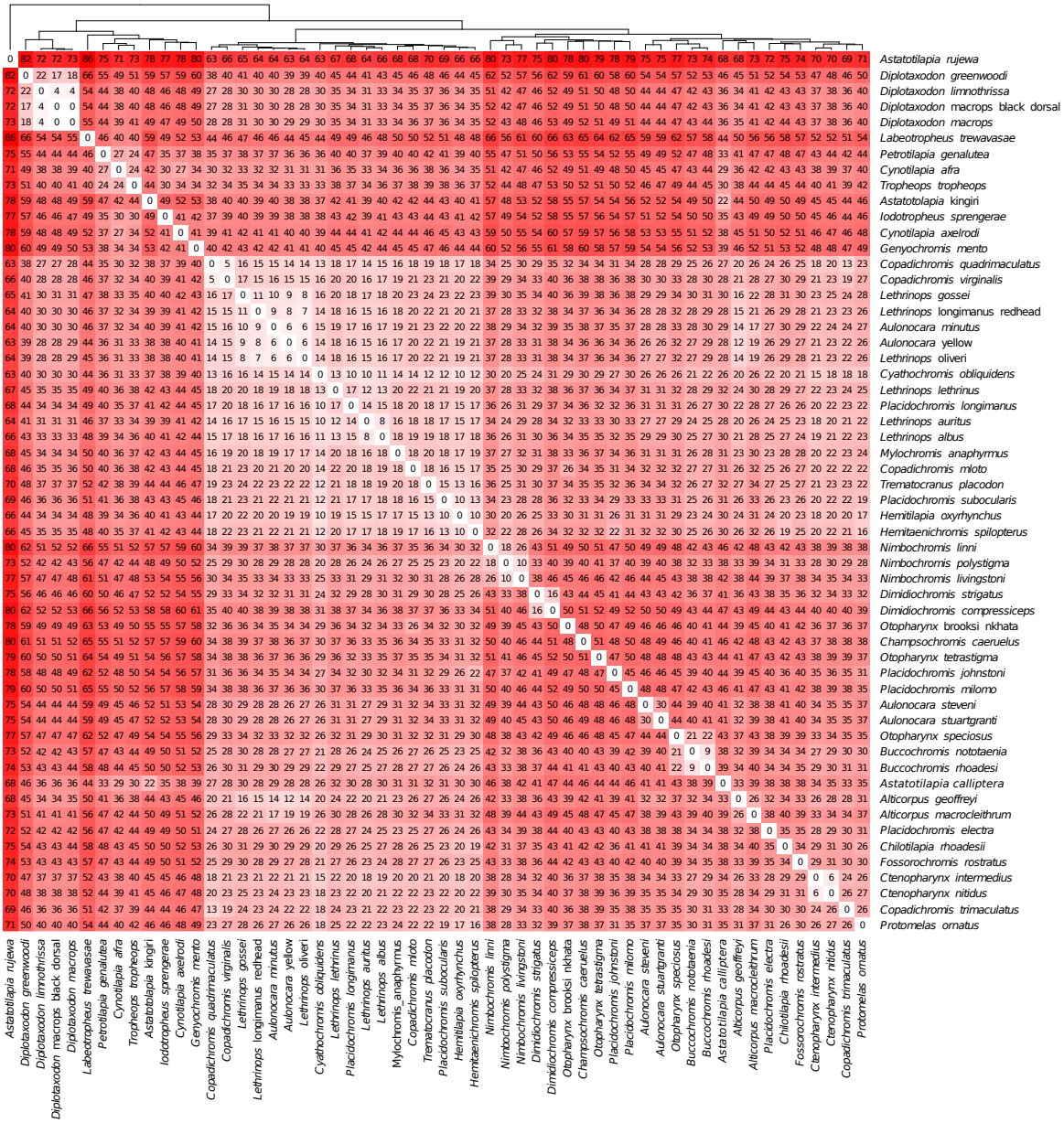

Fig. 5.3 **F_{ST} divergence between Lake Malawi species.** The tree (top) corresponds to simple hierarchical clustering of F_{ST} values by the `hclust` method in `R` [175]. F_{ST} values are given as percentages (%).

For the majority of Lake Malawi species, we have not sequenced enough individuals to study the genome-wide pattern of genetic differentiation and be able to identify loci that are under positive selection. Among exceptions to the above are *Placidochromis subocularis* and *Trematocranus placodon*. To test whether a genome scan for high F_{ST} outliers could be informative in this case, I calculated F_{ST} divergence between eight *P. subocularis* individuals and five *T. placodon* in non-overlapping sliding windows of 100 variants each.

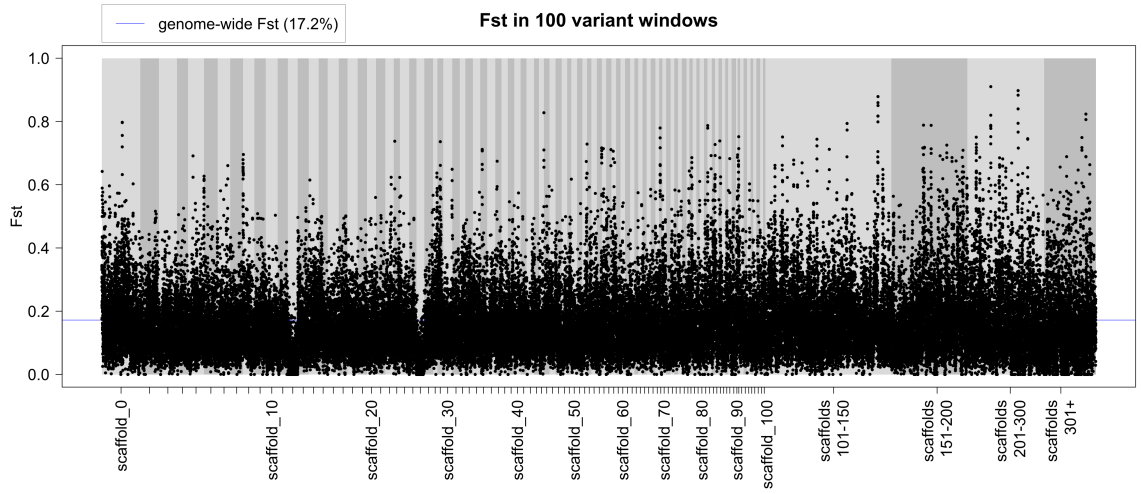

Fig. 5.4 **Genome-wide F_{ST} profile between *P. subocularis* and *T. placodon*.**

The results (Figure 5.4) suggest that neutral divergence between these two species, as measured by F_{ST}, is sufficiently low to allow for detecting selection. Also interesting is the pronounced dip in F_{ST} on scaffolds 12 and 26, which could suggest for example balancing selection, positive selection on the same ancestral variant in both species, or recent introgression between the two species. Nevertheless, it is worth keeping in mind that the numbers of individuals used in the analysis are likely too low to be able to reliably distinguish selection at specific loci from other causes of variation

Finally, to obtain an overview of the Lake Malawi dataset which includes both within- and between-species diversity I performed Principal Component Analysis (PCA). Figure 5.5 shows the position of each specimen along the three main principal components (three main axes of variation) which together explain 15.13% of the total genetic variation in the dataset. Within the space of this PCA projection, all specimens collected from the same species form closely knit clusters, as do some other groups comprised of several species: namely the mbuna, the genus *Diplotaxodon*, and all the deep water benthic specimens. Almost all shallow water sand-dwellers are located on a line along the second and third eigenvectors, except *Copadichromis virginalis* and

C. quadrimaculatus, both of which have evolved a more plankton feeding rather than bottom dwelling habit, and the shallow water members of the genus *Aulonocara*, who also are not typical sand-dwellers in that they prefer the intermediate habitat at the interface between sandy and rocky bottom areas.

The arrangement of species along a line in a PCA plot, as for the sand dwellers, is suggestive of patterns seen in modern human populations with varying degrees of admixture, such as Native Americans with European admixture. However, in this case we are considering separated species which may perhaps be subject to hybridisation and introgression at a low rate, rather than subpopulations of one species which have admixed due to population migrations in recent generations. I will return to evidence for hybridisation later in section 5.4.

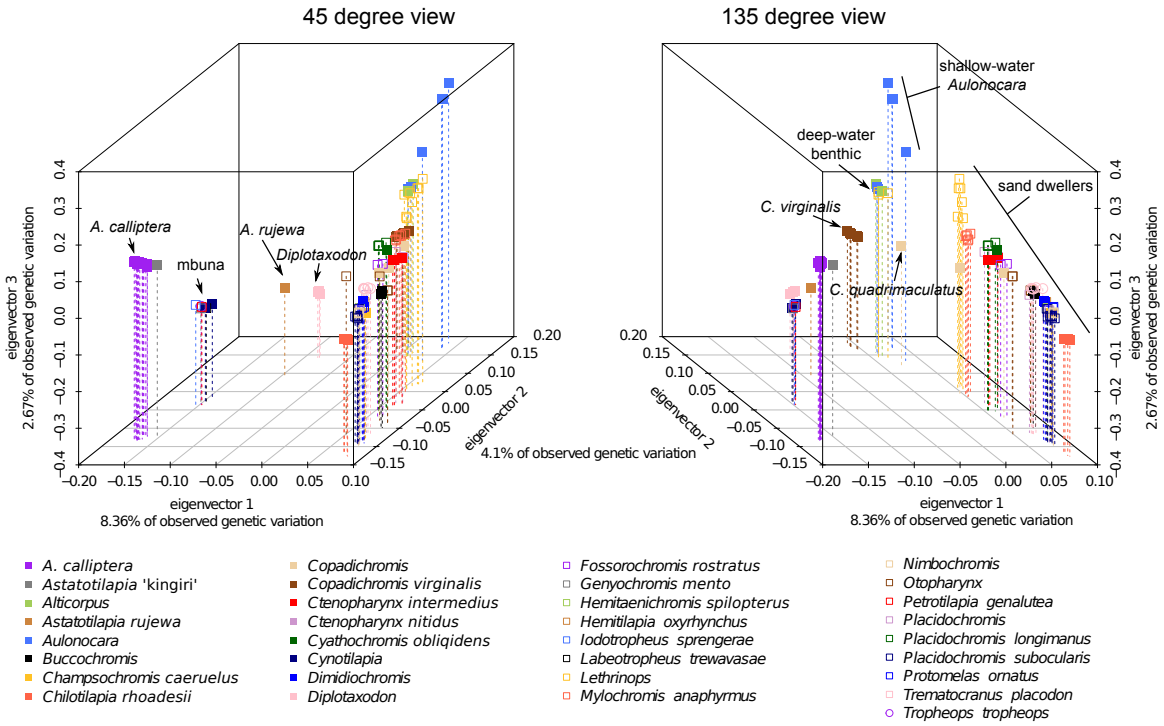

Fig. 5.5 **Principal Component Analysis of genetic variation in the Lake Malawi sample set.** Two views of the same 3D plot are shown, projecting the position of each sample along the first three principal components.

5.3 Phylogenetic relationships

The relative position of individuals in the PCA plot has given us some indication of the relationships between them. As a first step in trying to reconstruct the evolutionary

history or the Lake Malawi radiation, I used all the single nucleotide variants in the genome (except mtDNA) to build a 'consensus' phylogenetic tree with all the individuals in the sample set.

The tree (Figure 5.6), rooted using a whole genome alignment with the Lake Victoria *P. nyererei* (see section 3.4) reveals that:

1. All individuals assigned to the same species form monophyletic clades, strengthening the evidence for them being 'good' species and suggesting that, when averaging across all genomic loci, a phylogenetic approach to the study of Lake Malawi radiation can provide reliable information.
2. The whole genome signal confirms *Astatotilapia rujewa* as the closest available sister species to the Lake Malawi radiation, whereas the *Astatotilapia calliptera* clade is nested within the Lake Malawi radiation as a sister clade to the mbuna.
3. A number of genera, namely *Aulonocara*, *Copadichromis*, *Lethrinops*, *Otopharynx*, and *Placidochromis* are polyphyletic, revealing discrepancies between species relationships implied by traditional taxonomy and the whole genome DNA signal. For *Aulonocara* and *Lethrinops*, the split is between the shallow water (generally <40m) and deep water (generally >50m) species, with each group appearing to have different evolutionary histories. Samples from the genera *Otopharynx* and *Placidochromis* are polyphyletic within the sand dweller group, where taxonomists themselves (Eccles and Trewavas [129]) recognised they had difficulties in finding reliable phylogenetic signals in the morphology. The genus *Placidochromis* is defined by the absence of horizontal melanin pattern, presence of vertical bars, and absence of other defining characteristics [129]. The genus *Otopharynx* is defined on the basis of large melanin spots above the pectoral and/or anal fins, and again the absence of other derived characters [129]. The whole genome DNA phylogeny presented here suggests that these melanin patterns do not reliably indicate phyletic relationships and therefore are not good generic characters.
4. All *Copadichromis* samples, except *C. mloto*, are outside the sand dweller clade, suggesting that this genus of plankton feeders is indeed distinct from the other bottom dwelling fish.

Overall, the phylogeny presented here, while generally consistent with current nomenclature, suggests that for some genera a revision is required in the light of new molecular data. Interestingly, it suggests that ecological traits such as preferred water depth, previously unused in taxonomic revisions [129], may be useful phylogenetic characters.

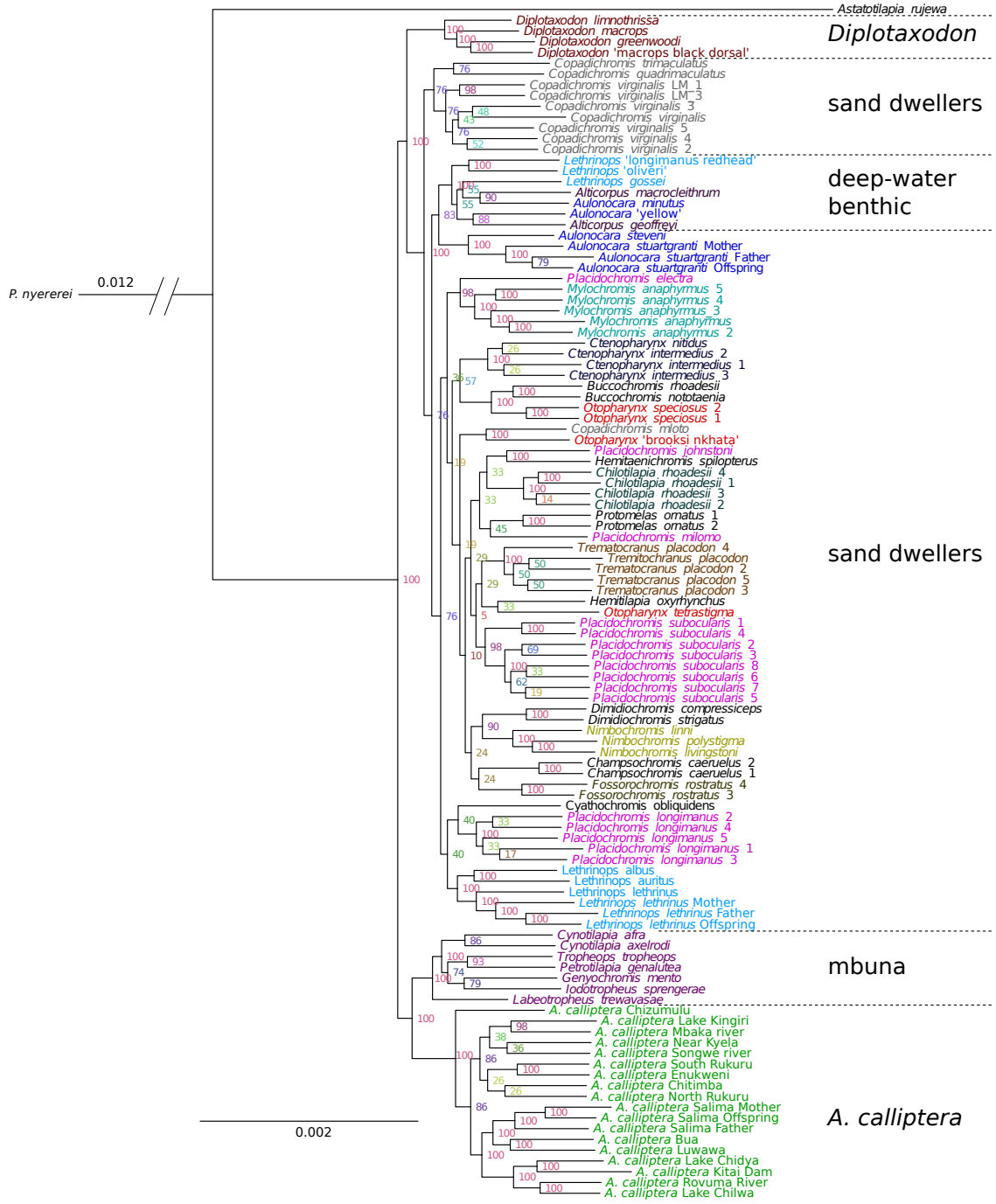

Fig. 5.6 **Lake Malawi whole genome maximum likelihood phylogeny.** Branch support values are based on 42 bootstrap replicates. Samples are coloured according to genera, except mbuna which are all displayed in a single colour.

Most previously published Lake Malawi phylogenies relied on mitochondrial DNA (e.g. ref [81, 86, 176]) or on a limited number (generally hundreds and up to two thousand) of nuclear variants generated with amplified fragment length polymorphisms (AFLP) [177] (e.g. ref [86, 176, 178, 179]).

Using the batch of Lake Malawi samples sequenced in Spring 2013 (see Table 3.2), I compared the mtDNA and whole genome consensus phylogenies. The results (Figure 5.7) show clearly that there are major differences between topologies of the two phylogenetic trees. The whole genome tree resolves species and known groups of species as monophyletic, although assumptions of the likelihood model are likely to be violated (this will be discussed in more detail later). In contrast, the major clades in the mtDNA tree do not correspond to any known or proposed groupings. Strong phylogenetic discordance between Lake Malawi mitochondrial and nuclear DNA phylogenies has been observed previously and interpreted as evidence for hybridisation and introgression between distinct taxa [86, 176].

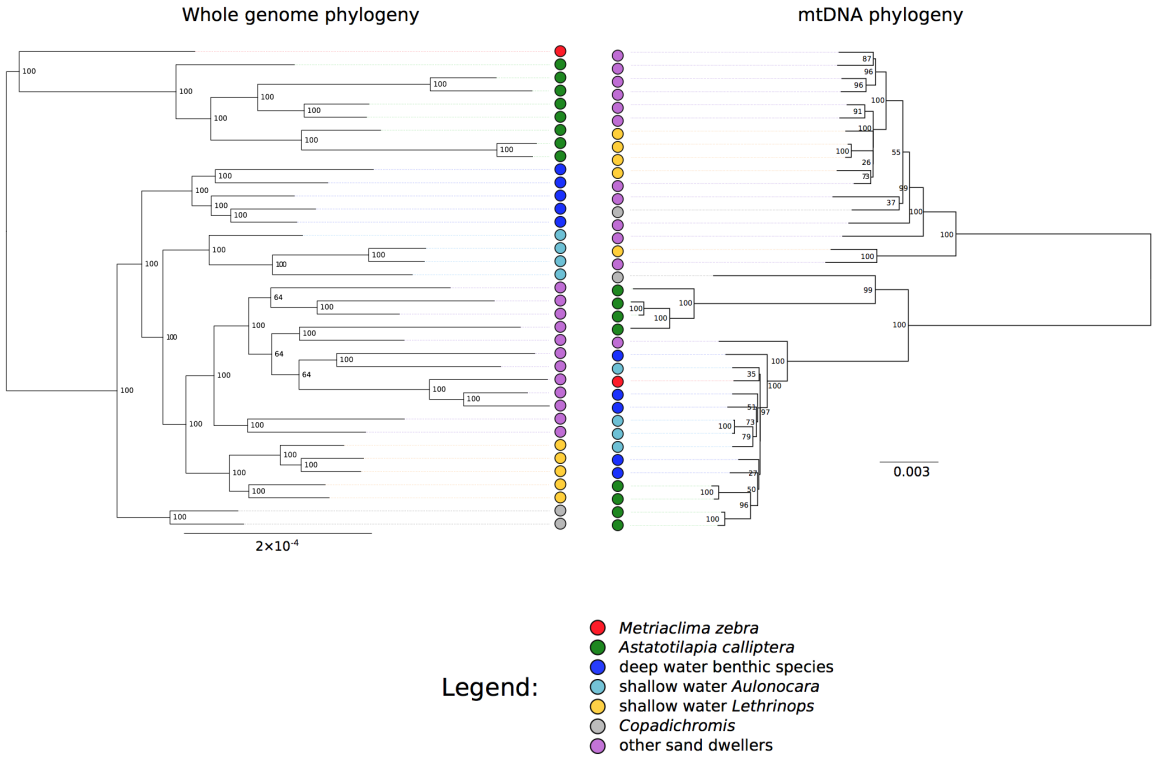

Fig. 5.7 **Contrast between whole genome and mtDNA phylogenies.** Branch support values are based on 281 bootstrap replicates for the whole genome phylogeny and 150 replicates for the mtDNA phylogeny. Both phylogenetic trees are rooted arbitrarily at their midpoints.

Next, I divided the genome into 1,460 regions, each comprising 5,000 segregating variants (mean region length 459kb, s.d. 141kb), and generated an independent phylogeny for each of these genomic regions. These *local phylogenies* are shown in Figure 5.8. Average bootstrap (over all branches) for 1,209 trees was above 40% and for 684 above 50% (Figure A.1), indicating that 5,000 variants were sufficient to provide an informative phylogenetic signal for the majority of the trees.

I found that, even if branch lengths are ignored, all trees differ from each other in their topologies (i.e. no two trees imply the same relationships between the species), implying pervasive effects of incomplete lineage sorting and/or introgression (see section 1.2). It is also notable that all 1,460 trees differ in their topology from the overall whole-genome consensus tree. However, given that both the mean and the standard deviation in the time to coalescence are $2N_e$ generations (90,000 years with N_e=15,000 and average generation time of three years), it is not unexpected that lineages reflecting within-species variation tend not to coalesce within the duration of the species.

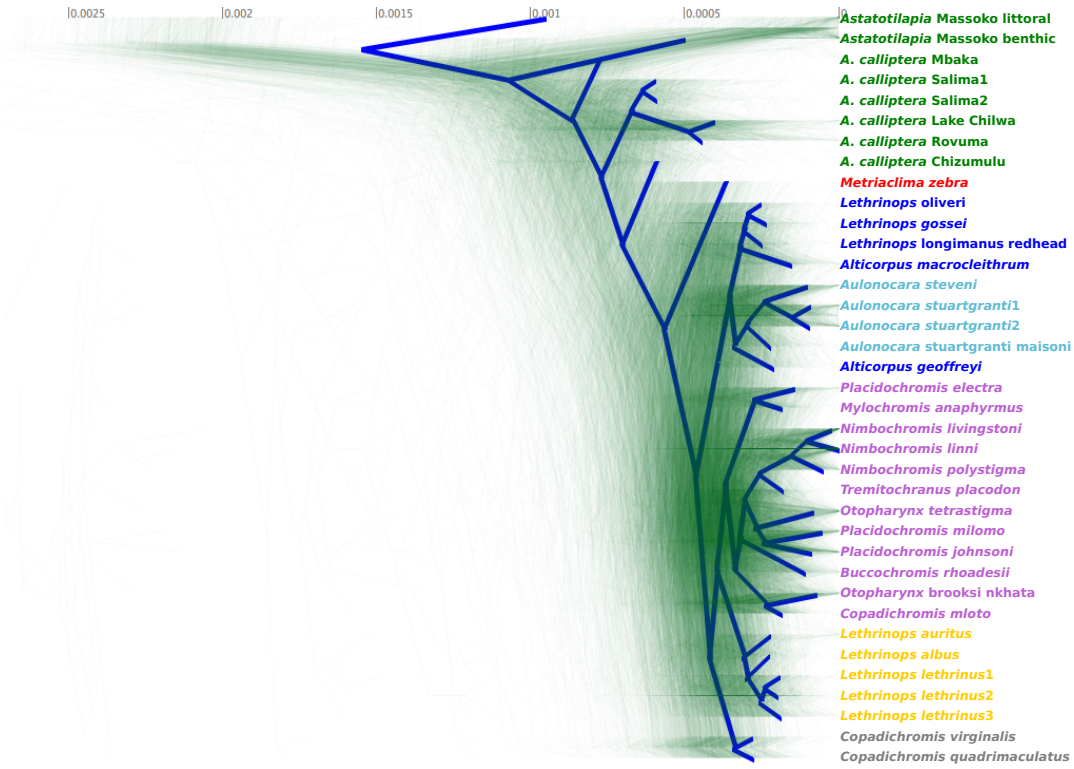

Fig. 5.8 **Variation in phylogenies between 1,460 regions along the genome.** Each tree is based on 5,000 consecutive variants. The tree were drawn using the `DensiTree v2.1` software [180] and the blue consensus tree has been obtained using the 'root canal' function of DensiTree.

It is worth noting here that the large amount of incomplete lineage sorting (ILS) described above means that the assumptions of the likelihood model used to build the whole genome phylogeny are likely to be substantially violated. The model assumes that the whole sequence has been generated by a single bifurcating tree; therefore, every instance of ILS is assumed to be the result of a repeated mutation at the same site, a highly improbable event. Therefore, future phylogenetic analysis in this system will need to also consider other approaches, such as the Neighbour-Joining method [181].

The Robinson-Foulds (RF) dissimilarity metric offers an objective procedure for comparing phylogenetic trees [182]. I used the `RF.dist` function implemented in the `phangorn` [183] package for `R` to calculate this metric for a random sample of 1,460 pairs of local phylogenies, and also to calculate distances between local phylogenies and the mtDNA tree. Figure 5.9A compares the two distributions of distances, showing clearly that the mtDNA tends to be more dissimilar (i.e. its topology tends to be more different from the local phylogenies based on nuclear DNA than they are from each other). I also calculated the RF metric between the mtDNA and whole genome phylogenies (distance 60). Only four out of the 1,460 local trees based on nuclear DNA have RF distance from the whole genome phylogeny $\geq$60.

From the point of view of population genetics, the main difference between mtDNA and nuclear DNA is that mtDNA is inherited only through the female lineage and is haploid. Assuming that there are equal numbers of reproducing males and females, this means that the effective population size for mtDNA is four times lower that for nuclear DNA (for a detailed discussion, including exceptions, see [184]). The lower N_e results in a different distribution of coalescence times (Figure 5.9) so mtDNA alleles tend to coalesce in different ancestral species than nuclear DNA alleles.

Another factor contributing to the mtDNA phylogeny being an outlier is the lack of recombination in the mitochondria; the mtDNA tree is just a single genealogy drawn from the distribution of many possible genealogies shaped by the population history and speciation events. On the other hand, a tree from even just a ~500kb segment of the genome is likely to be an average of several genealogies separated by ancestral recombination events.

Finally, mtDNA may be more likely to introgress between closely related species than nuclear DNA. A large number of studies have reported extensive mtDNA introgression with little or no evidence of introgression in the nuclear genome, e.g in fish [185], amphibians [186], birds [187], mammals [188], and fruit flies [189]. The factors that can explain these discrepancies are reviewed in ref [184]. While some of the cited studies rely on just a few microsatellite nuclear markers [186, 187], Good *et al.* [188] use

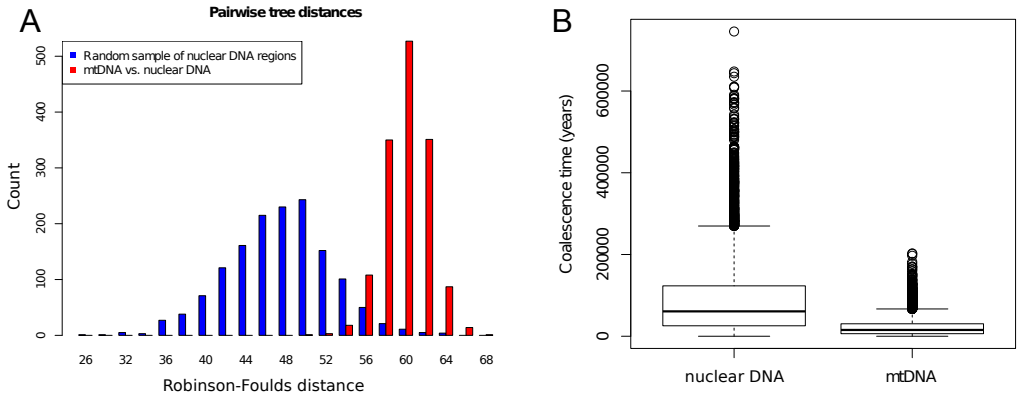

Fig. 5.9 **Difference between local phylogenies based on nuclear DNA and the mtDNA phylogeny. (A)** Robinson-Foulds distances. **(B)** Simulated coalescence times for nuclear DNA and mtDNA with exponential distributions, assuming an ideal population, N_e=15,000 and that the average generation time is three years. Coalescence times for nuclear DNA are then distributed as ~exp($\frac{1}{90{,}000}$) and mitochondrial ~exp($\frac{1}{22{,}500}$). I simulated 5,000 observations from each distribution.

sequence data from 10,500 genes to convincingly rule out "all but very minor levels of interspecific gene flow" despite complete mtDNA replacement between the two rodent species in their study. Therefore, in order to understand the importance and extent of introgression in the Lake Malawi radiation it is necessary to re-evaluate the evidence from a genome-wide perspective, rather than rely on mtDNA evidence.

5.4 Exploring evidence for interspecific introgression

5.4.1 ABBA-BABA tests

The ABBA-BABA statistic (also known as the *D statistic* or *Patterson's D*) tests for an excess of shared derived alleles between one of two populations (P_1 and P_2) and their outgroup (P_3) [190, 191]. I defined derived alleles with respect to the Lake Victoria species *P. nyererei*. Under the null hypothesis of no differential gene flow from P_3, the two populations P_1 and P_2 are expected to share derived alleles with P_3 equally often, while introgression from P_3 to P_1 or from P_3 to P_2 would result in excess sharing, as illustrated in Figure A.2.

I used the ABBA-BABA statistic to test if the phenotypic similarity between the shallow and deep water species of the genus *Lethrinops* could be due to introgression from shallow water *Lethrinops* into a deep water relative of the genus *Alticorpus*. In the

Lake Malawi whole genome phylogeny (Figure 5.6), shallow and deep water *Lethrinops* form two distinct groups, but their anatomy is sufficiently similar for them to be assigned to the same genus by Eccles and Trewavas [129]. I therefore tested for a signature of gene flow from the shallow water *L. lethrinus* into the deep water species. If that were the case, I expected to find an excess of the ABBA pattern, as illustrated by the test design shown in Figure 5.10.

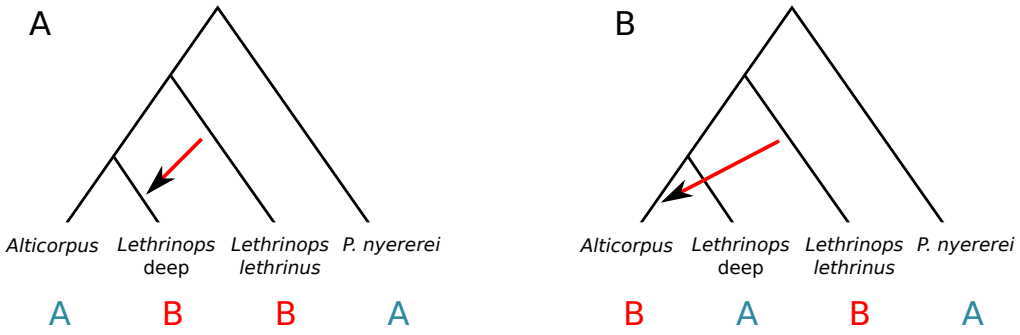

Fig. 5.10 **The ABBA-BABA test for introgression from shallow to deep water species of the genus *Lethrinops*.** The ancestral allele is denoted by the A character and the derived allele by the B character.

Overall, I did not find any excess of shared derived alleles between *L. lethrinus* and deep water *Lethrinops* when compared with the deep water genus *Alticorpus* (Patterson's D=0.37%; 1.28 SD from 0% or P=0.101). Therefore, significant levels of gene flow specifically into the deep water *Lethrinops* species can be ruled out. To explore whether smaller regions of the genome might have introgressed, perhaps due to selection following low levels of hybridisation, I also calculated the D statistic for non-overlapping genomic regions of 100 informative variants each (informative in the sense that they change the numerator of the D statistic) (Figure 5.11).

The regions for local D statistic calculation were on average 137kb long, but with large variation (from 4kb to 911kb; s.d. 63kb). I found several outlier regions with D statistic more than ±4.5 s.d. away from the mean. If introgression happened, these would be the best candidates for introgressed regions. However, it is important to note that there are outlier loci in both directions, i.e. suggesting introgression both into deep water *Lethrinops* and into *Alticorpus*. While this is not impossible, it is advisable to be cautious as the locally calculated D statistic has been shown by simulations to have large variance [192].

For all of the above ABBA-BABA analysis I had access only to the samples sequenced in Spring 2013, so three *L. lethrinus*, three deep water *Lethrinops* and two

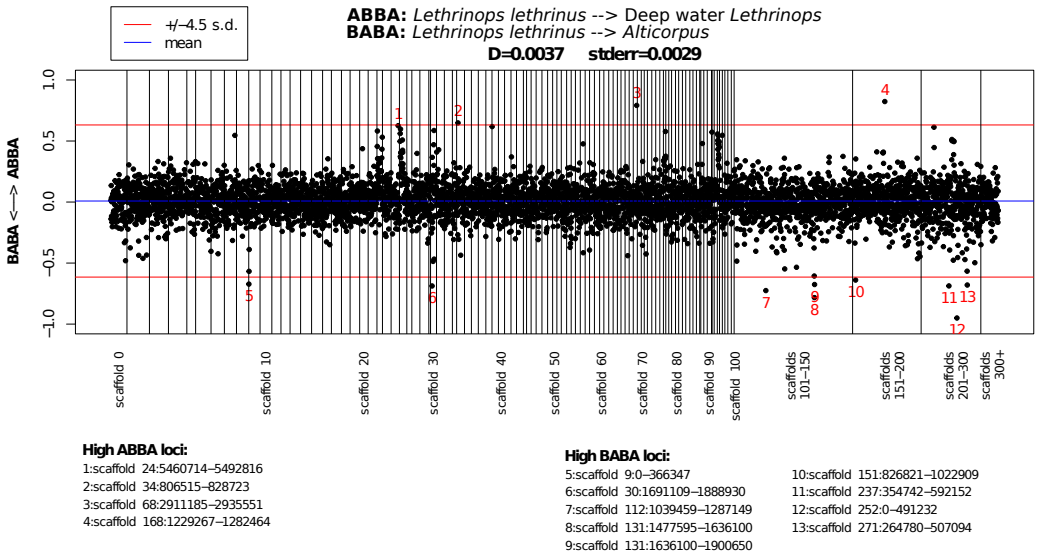

Fig. 5.11 **The results of the ABBA-BABA test for introgression from shallow water to deep water species of the genus *Lethrinops*.** Each point corresponds to the value of D statistic calculated over a window of 100 informative variants. Outlier regions are listed below the plot.

Alticorpus individuals. This was sufficient when calculating the D statistic over the entire genome. However, reducing neutral variance in the local introgression analysis will require adding more samples per group and using new statistics specifically designed to locate introgressed loci such as f_d [192] and f_{dM} (Methods - section 5.5).

5.4.2 Chromopainter and fineSTRUCTURE

I used the `Chromopainter` program [193] to 'paint' the chromosomes of each individual based on haplotype similarity with chunks of DNA from the other individuals. Each chromosome was conceptually segmented into regions, bounded by ancestral recombination sites, each of which is matched with another chromosome which is assigned as the 'donor' or the local 'closest relative'. This is done statistically, resulting in a measure of co-ancestry to each other chromosome in the data set. The results of this 'chromosome painting' for all Lake Malawi samples are summarised in the co-ancestry matrix shown in Figure 5.12. Individuals have been clustered by the `fineStructure` software [193] based on the amounts of co-ancestry they share. Overall, the co-ancestry matrix provides a view of Lake Malawi species relationships based on haplotype relationships, which is complementary to the phylogenetic analyses and the F_{ST} distances presented previously.

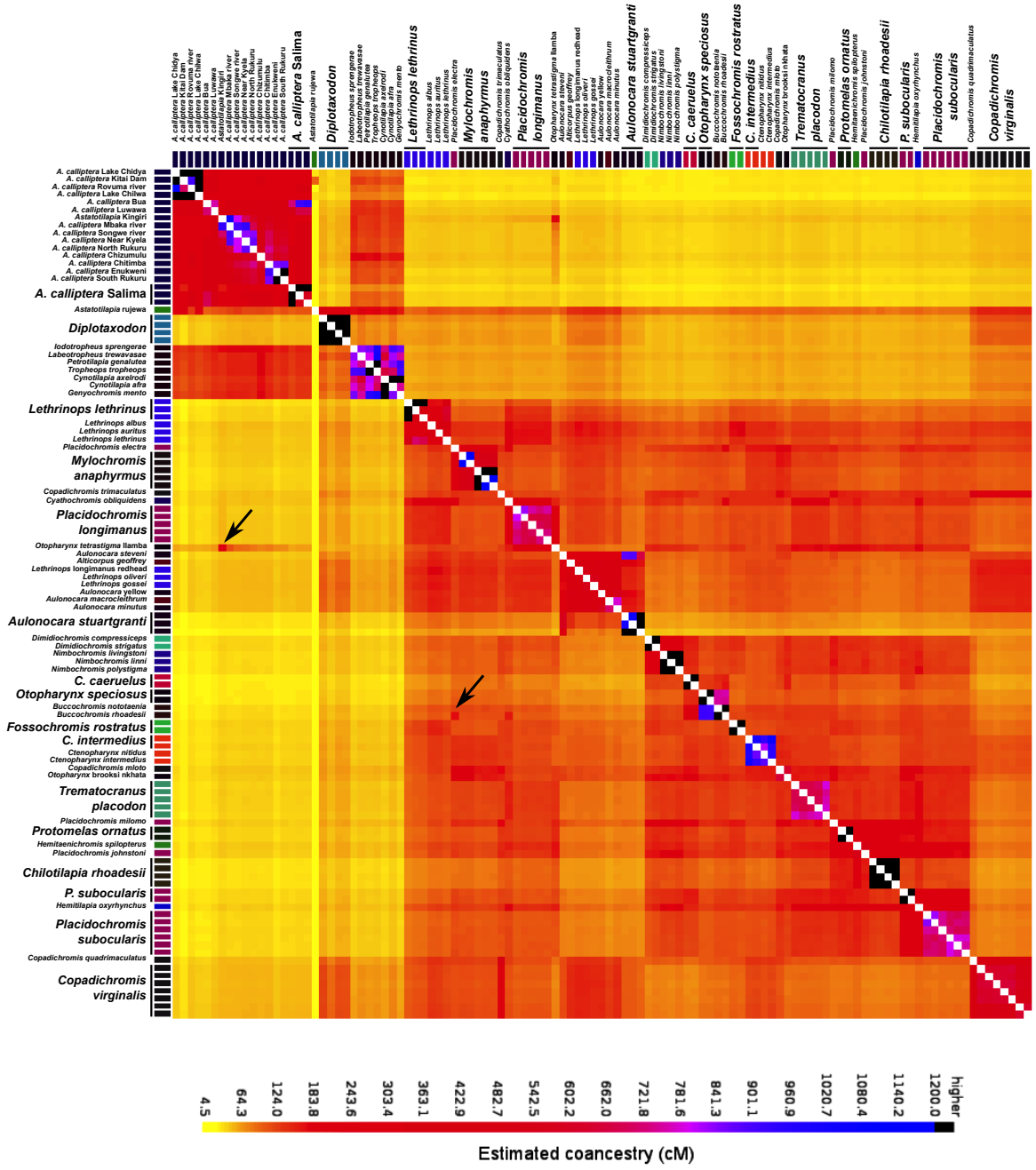

Fig. 5.12 **Co-ancestry between Lake Malawi samples measured by the Chromopainter software.** Each row corresponds to a 'recipient' and each column to a 'donor'. Thus the values indicate the total length (in cM) along which a 'donor' haplotype was inferred to be the closest relative of a 'recipient' haplotype. Clustering was done by the fineStructure software [193]. Examples of possible introgression are indicated by arrows.

Sharing of chunks of chromosomes between otherwise distantly related species would be an indication of recent admixture or introgression. The Chromopainter co-ancestry matrix provides several such hints. For example, there is unexpectedly high co-ancestry between the *Otopharynx tetrastigma* from Lake Ilamba and the *Astatotilapia* from Lake Kingiri (Figure 5.12), suggesting the *O. tetrastigma* and *Astatotilapia* may hybridise in Ilamba (we have not yet sequenced an Ilamba *Astatotilapia*, so we see the strongest signal in the closely related Kingiri sample). Other, although not as strong, hints of introgression can be seen within the shallow water sand dweller group. For example, *Buccochromis rhoadesii* has unexpectedly high co-ancestries with *Placidochromis electra* and with *Cyathochromis obliquidens*.

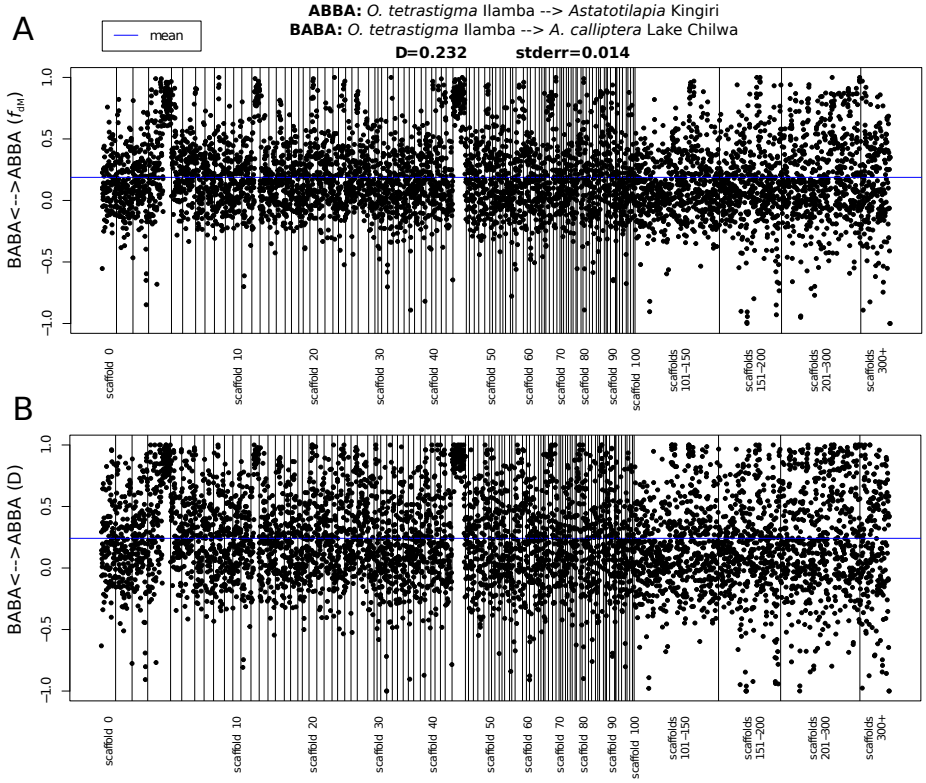

Fig. 5.13 ***O. tetrastigma* Ilamba introgression into crater lake *Astatotilapia*.** **(A)** Point correspond to the f_{dM} statistic calculated over windows of 50 informative variants. **(B)** The D statistic calculated over windows of 50 informative variants.

I followed up on the hypotheses of introgression generated by the Chromopainter software by calculating the ABBA-BABA statistics. First, I found a large excess of shared derived alleles between *O. tetrastigma* from Ilamba and *Astatotilapia* Kingiri, when compared with *A. calliptera* from Lake Chilwa (Patterson's D=23.2%; 16.6 SD from 0% or P<3.4×10^{-62}). The proportion of admixture f [190] was estimated at 29.8±1.7%. Next, I calculated local D and f_{dM} statistics for non-overlapping genomic

regions of 50 informative variants each (Figure 5.13). The results confirm that the f_{dM} statistics has lower variance (s.d. = 0.31 against 0.37 for D) and suggest that scaffolds 3, 12, and 44 harbour regions with especially strong signatures of introgression.

Next, I tested for possible introgression from *Placidochromis electra* into *Buccochromis rhoadesii*. I found a small excess of shared derived alleles between these two species, when compared to sharing between *P. electra* and *B. notoaenia* (Patterson's D=1.31%; 2.52 SD from 0% or P=0.006). The proportion of admixture f was estimated at 1.3±0.9%. Therefore, the genome-wide statistics suggest that there has been a small amount of introgression from *P. electra* into *B. rhoadesii*. However, unlike in the previous case of *O. tetrastigma* hybridisation in Ilamba, local D and f_{dM} statistics do not provide clear pointers to specific introgressed regions (Figure 5.14).

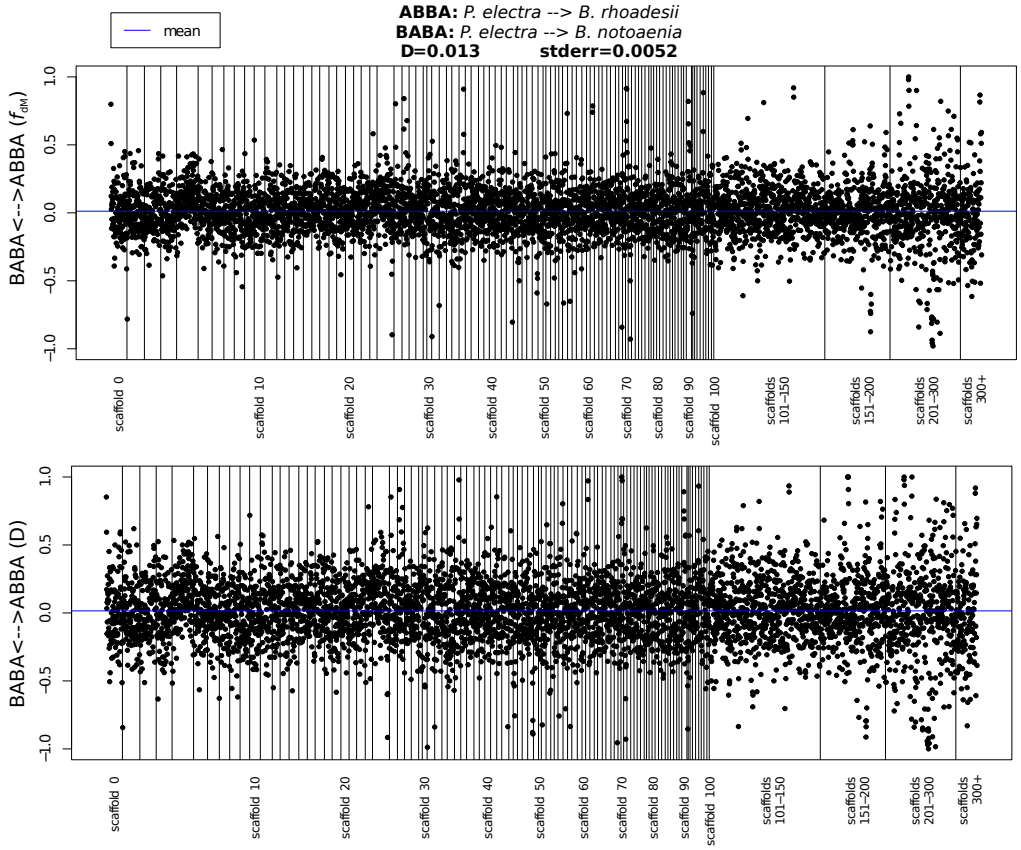

Fig. 5.14 ***Placidochromis electra* introgression into *Buccochromis rhoadesii*.** **(A)** Point correspond to the f_{dM} statistic calculated over windows of 50 informative variants. **(B)** The D statistic calculated over windows of 50 informative variants.

5.5 Methods

Basic statistics

Basic statistics (number of variants called, numbers of heterozygous sites) were obtained using the `bcftools v1.1 stats` tool, with the `-s -` option to include all samples. Results from the 'Per-sample counts' (PSC) lines were then plotted using the `R` software environment [175] .

Principal Component Analysis

SNP variants (no indels) with minor allele frequency >= 0.05 were selected using `vcftools v0.1.12b` options `--remove-indels --maf 0.05` and exported in `PLINK` format [194]. The variants were LD-pruned to obtain a set of variants in approximate linkage equilibrium (unlinked sites) using the `--indep-pairwise 50 5 0.2` option in `PLINK v1.0.7`. Principal Component Analysis on the resulting set of variants was performed using the `smartpca` program from the `eigensoft v5.0.1` software package [195] with default parameters.

The 3D plots were generated with the using the `scatterplot3d` [196] package for `R`.

Genome-wide F_{ST} calculations

All SNP variants were selected and LD-pruned to obtain a set of variants in approximate linkage equilibrium (unlinked sites) using the `--indep-pairwise 50 5 0.2` option in `PLINK v1.0.7`. Next I used the `smartpca` program from the `eigensoft v5.0.1` software package [195] with adding the parameter `outliersigmathresh: 11` to prevent the removal of outliers. The `smartpca` program calculates genome-wide F_{ST} for all pairs of populations specified by the sixth column in the `.pedind` file using the Hudson estimator, as defined by Bhatia, Patterson *et al.* [197] in equation 10, and using 'ratio of averages' to combine estimates of F_{ST} across multiple variants, as recommended in their manuscript [197].

Local F_{ST} calculations

For sliding-widow based analysis, I used my own implementation of the Bhatia, Patterson *et al.* [197] F_{ST} calculation in the C++ program `evo` (available from `https://github.com/millanek/evo` with the `fst --vcf` option).

Phylogenetic trees

For the whole genome maximum likelihood phylogeny in Figure 5.6, I generated consensus genome sequences using the `bcftools v1.2 consensus` tool. For each sample, I selected the sequence of one haplotype (as assigned by `beagle` haplotype phasing - see section 3.2.4) by using the `--haplotype=1` option in `bcftools`. All scaffolds except the mtDNA sequence (scaffolds 747, 2036) were concatenated into a single sequence and phylogenetic trees then inferred using `RAxML v7.7.8` [198] under the GTRGAMMA model (General Time Reversible model of nucleotide substitution with the Γ model of rate heterogeneity). The maximum likelihood tree was obtained as the best out of five alternative runs on distinct starting maximum parsimony trees (using the `-N 5` option). Forty two bootstrap replicates were obtained using `RAxML's` rapid bootstrapping algorithm [199]. It was my intention to run more bootstrap replicates, enough to satisfy `RAxML -N autoFC` frequency-based bootstrap stopping criterion, but this has proven computationally infeasible on a dataset of this size (obtaining the 42 replicates required thousands of hours of CPU time). Still 42 replicates provide a reasonable indication of bootstrap support for the maximum likelihood tree. Bipartition bootstrap support was drawn on the maximum likelihood tree using `RAxML -f b` option.

For all other phylogenies using samples from Spring 2013 (Figures 5.7, 5.8), I generated consensus sequences using unphased genotype data with my own C++ program `evo` (available from `https://github.com/millanek/evo`) using the `getWGSeq --whole-genome` options for the whole genome data, `getWGSeq --mtDNA` for mitochondrial sequences, and `getWGSeq --split 5000` for local phylogenies. Heterozygous variants were represented by IUPAC codes. The phylogenies were then built in the same way as described above, with one difference: the number of bootstrap replicates was always determined by the `RAxML -N autoFC` frequency-based bootstrap stopping criterion.

Chromopainter and fineSTRUCTURE

Singleton SNPs were excluded using `bcftools-1.1 -c 2:minor` option, before exporting the remaining variants in PLINK format [194]. The `chromopainter v0.0.4` software [193] was then run for 150 largest genomic scaffolds. Briefly, I created a uniform recombination map using the `makeuniformrecfile.pl` script, then estimated the effective population size (N_e) for a subsample of 20 individuals using the chromopainter inbuilt expectation-maximization procedure [193], averaged

over the 20 N_e values using the provided `neaverage.pl` script. The chromopainter program was then run for each scaffold independently, with the `-a 0 0` option to run all individuals against all others. Results for individual scaffolds were combined using the `chromocombine` tool before running `fineSTRUCTURE v0.0.5` with 1,000,000 burn in iterations, and 200,000 sample iterations, recording a sample every 1,000 iterations (options `-x 1000000 -y 200000 -z 1000`). Finally, the sample relationship tree was built with fineSTRUCTURE using the `-m T` option and 20,000 iterations.

Patterson's D (ABBA-BABA) and related statistics

I calculated the D statistics using equation S15.2 of Green et al. [190], allowing me to use allele-frequency information from all benthic and littoral individuals. I also estimated *f*, the admixture fraction following Green *et al.* equation S18.5, and calculated the standard error for both estimates by a weighted block jackknife, using blocks of 5,000 informative variants (i.e. variants with ABBA or BABA patterns).

I also calculated a version of the f_d statistic designed by Martin *et al.* specifically to detect introgressed loci [192, equation 6]. One of the limitations of the f_d statistic is that it is not symmetric; it is distributed on the interval (-Infinity,1]. Therefore, I define a closely related statistic which I call f_{dM}. Compared with the f_d statistic, f_{dM} has the advantages that it is bounded on a the interval [-1,1], and under the null hypothesis of no introgression is symmetrically distributed around zero.

Following the notation from Martin et al. [192], I consider three populations and an outgroup with the relationship (((P$_1$, P$_2$), P$_3$), O). Let:

$$S(P_1; P_2; P_3; O) = \sum_i ((1 - p_{i1})p_{i2}p_{i3}(1 - p_{i4})) - \sum_i (p_{i1}(1 - p_{i2})p_{i3}(1 - p_{i4})) \quad (5.2)$$

where p_{ij} is the frequency of the derived allele at site i in population j.

f_{dM} is then defined as follows:

$$f_{dM} = \begin{cases} \dfrac{S(P1; P2; P3; O)}{S(P1, PD, PD, O)} = f_d, & \text{if } p_{i2} \geq p_{i1} \\ \dfrac{S(P1; P2; P3; O)}{-S(PD, P2, PD, O)}, & \text{otherwise} \end{cases}$$

where P_D is the population (either P$_1$ or P$_3$) that has the higher frequency of the derived allele. For a detailed discussion of the f_d statistic see Martin et al. [192].

The D and f statistics were calculated genome-wide and D and f_{dM} also in non-overlapping windows of 50 or 100 informative variants each, as indicated in the main text. To add ancestral allele information (i.e. the outgroup variants) to the Lake Malawi set VCF file, I used whole genome alignment between *M. zebra* and *P. nyererei*, as described in section 3.4.